CS 70 Discrete Mathematics and Probability Theory
Spring 2022 Course Notes **Note 20**

Prediction!

Prediction is one of the most compelling applications of probability, particularly in computer science.

As an example, suppose you know the height of all students at UC Berkeley, and a Berkeley student is chosen uniformly at random. What would be the best prediction for that student's height that minimizes the error of your prediction? It would make sense to guess the average height, as it lies in the "middle" of all the heights.

More formally, prediction is the problem of making an estimate for a random variable from available information; for this note, that information is a probability distribution or joint distribution. The goal of prediction is to minimize the error (sometimes called loss) in the prediction.

More formally, we consider the *mean squared error*. That is, for a random variable X with a known distribution, we define prediction as giving an estimate $\hat{x}$[1] which minimizes the following expression:

$$\mathbb{E}[(X-\hat{x})^2].$$

The solution to this prediction problem is $\hat{x} = \mathbb{E}[X]$!

1 The parabola

Mathematically, we revisit those childhood days of understanding the parabola. Recall that for a parabola defined by $f(x) = ax^2 + bx + c$, the vertex of the parabola is at $x = -\frac{b}{2a}$. Long ago, you derived this by "completing the square".

Exercise: Try completing the square at home! Its really a lovely idea.

One can also use calculus and set the derivative, $f'(x) = 2ax + b$, to zero, and solve for x to find a critical point. It is a minimum when a is positive (opens upward graphically) and a maximum when a is negative.

2 Expected value is optimal!

Recall, we wish to find the value for $\hat{x}$ which minimizes $\mathbb{E}[(X-\hat{x})^2]$. Expanding gives

$$\begin{aligned}\mathbb{E}[(X-\hat{x})^2] &= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\,\mathbb{E}[\hat{x}] + \mathbb{E}[\hat{x}]^2 \\ &= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\hat{x} + \hat{x}^2\end{aligned}$$

The second line follows from the fact that $\hat{x}$ is a constant with respect to the expectation over X. We see that this is a parabola of the form $a\hat{x}^2 + b\hat{x} + c$ where $a = 1, b = -2\,\mathbb{E}[X]$ and $c = \mathbb{E}[X^2]$. (Since X is a known

[1] Throughout the note, we will use the notation $\hat{x}$ for predictions, to distinguish predictions from random variables and their outcomes.

distribution, b and c are constant.) Thus, the minimum is at $\hat{x} = -\frac{-2\mathbb{E}[X]}{2} = \mathbb{E}[X]$. Middle school is cool. [2]

An important observation is that the minimum value of the mean squared error of this estimator is exactly the variance of X, i.e., $\mathbb{E}[(X - \mathbb{E}[X])^2]$.

An important conceptual frame in estimation, what is the value of the best estimate and what is the error. In this setting, the estimator for X with the lowest mean squared error is $\mathbb{E}[X]$, and this estimate results in a mean squared error of $\mathrm{Var}(X)$. Do take note of this!

3 Joint Distributions: conditional expectation.

Recall the following definition that we presented previously.

Definition 20.1. *The joint distribution for two discrete random variables X and Y is the collection of values $\{((a,b), \mathbb{P}[X = a, Y = b]) : a \in \mathscr{A}, b \in \mathscr{B}\}$, where $\mathscr{A}$ is the set of all possible values taken by X and $\mathscr{B}$ is the set of all possible values taken by Y.*

When given a joint distribution for X and Y, the distribution $\mathbb{P}[X = a]$ for X is called the *marginal distribution* for X, and can be found by "summing" over the values of Y. That is,

$$\mathbb{P}[X = a] = \sum_{b \in \mathscr{B}} \mathbb{P}[X = a, Y = b].$$

The marginal distribution for Y is analogous, as is the notion of a joint distribution for any number of random variables.

From a joint distribution, we can compute the *conditional probability* of $X = x$ given that $Y = y$ from the definition of conditional probability as follows:

$$Pr[X = x | Y = y] = \frac{Pr[X = x, Y = y]}{Pr[Y = y]}.$$

The *conditional expectation* of X given $Y = y$ is defined naturally as follows:

$$E[X | Y = y] = \sum_{x \in \mathscr{A}} x Pr[X = x | Y = y].$$

That is, $E[X|Y = y]$ is simply the expectation of X given that $Y = y$. This is useful for prediction in the same sense that expectation is useful as we will return to later in this note.

3.1 Iterated Expectation and Wald's identity.

Before returning to prediction, we discuss *the law of iterated expectations* which is

$$E[X] = E(E[X|Y]) = \sum_{y \in \mathscr{B}} E[X|Y = y] Pr[Y = y]. \tag{1}$$

This simple concept can be quite useful. For example, consider choosing an integer N at random and forming a random variable $Y = X_1 + \cdots X_N$ where the X_i's are identical and independently distributed. Note

[2] The joke is that one learns about the parabola in middle school. Unfortunately, middle school is anything but fun. Also, unfortunately, having to explain a joke perhaps means it is a poor one.

the *number* of terms is a random variable here! We wish to compute $E[Y]$ and assuming that the random variables X_i is independent of the value of N.

$$\begin{aligned}
E[Y] &= E[E[Y|N]] \\
&= \sum_n E[Y|N=n]Pr[N=n] \\
&= \sum_n nE[X_1]Pr[N=n] \\
&= E[X_1]\sum_n nPr[N=n] \\
&= E[X_1]E[N]
\end{aligned}$$

The third line follows from $Y = X_1 + \cdots + X_n$ and the fact that the X_i's are identically distributed and are independent of the value of N. Thus, we have $E[Y] = E[X_1]E[N]$ which is the basic form of *Wald's identity.*

This can be useful for modeling the total time to serve customers in a time interval, where we have "Poisson" arrivals, and each customer's service time is from the same distribution. That is, we have a Poisson random variable, $N \sim P(\lambda)$, that determines the number of customers and each X_i is a random variable that corresponds to the time needed to serve customer i.

To conclude, we note that the law of iterated expectations is sometimes called the tower rule as one can extend the concept to more than two random variables, e.g., $E[X] = E[E[E[X|Y,Z]]]$, where the outer expectations are over the values of Y and Z analogous to 1.

4 Minimum Mean Square Estimate (MMSE): an example.

Returning to prediction: one predicts given more information just than one's prior expectations or (in our words) distributions to predict a value. We do this constantly in our behavior, e.g., when will a traffic light turn red after we see it turn yellow?

A simpler example is to predict the weight of someone from their height. That is, our predictor for someone's weight is based on their height h, so we will write the predictor as a function $\hat{w}(h)$ of h rather than as a constant. Here we have two pieces of information to base our prediction on. The first is the joint distribution of the weights, W, and heights, H. That is, we have $\mathbb{P}[W = w, H = h]$ for any weight w and height h. The second is the individual's height itself, h.

Exercise: What estimate should you use for W to minimize mean squared error if you do not know the height?

For the known value of h, we wish to find $\hat{w}(h)$ that minimizes the mean squared error which is:

$$\mathbb{E}[(W - \hat{w}(h))^2 | H = h].$$

Recall, that conditioning on $H = h$ yields a distribution on W which we denote as $\mathbb{P}[W = w|H = h]$, which in turn has a conditional expectation, denoted by $\mathbb{E}[W|H = h]$. This is just the expectation of W in the sample space corresponding to the event $H = h$, and thus by the previous section is the best estimate for W with regards to minimizing the mean squared error. That is, the best predictor for each height h is $\hat{w}(h) = \mathbb{E}[W|H = h]$.

We can also see that the mean squared error $\mathbb{E}[(W-\hat{w}(H))^2]$ is minimized with respect to the entire sample space for this predictor. In particular, recall that by law of total expectation,

$$\mathbb{E}[(W-\hat{w}(H))^2] = \sum_h \mathbb{P}[H=h] \times \mathbb{E}[(W-\hat{w}(h))^2 | H=h].$$

Since we can make a different prediction for each height we observe, to minimize this sum, it is enough to minimize each $\mathbb{E}[(W-\hat{w}(h))^2|H=h]$ on its own.

So the prediction $\hat{w}(h) = \mathbb{E}[W|H=h]$ also minimizes the expected squared error over the entire joint distribution. This is subtle, but should be more clear after we discuss linear regression.

4.1 Application and warning.

An example of doing this, informally, might be grouping people into buckets according to height, and predicting the weight of a random person in a bucket as the "average" or expectation of the people in the bucket.

The process is a way of approximating the joint distribution of height and weight. Statisticians are careful when doing so to distinguish the actual mean (or the conditioned expectation) from the real underlying mean (or conditioned expectation) for example by denoting the mean by $\tilde{\mu}$ and the real and unknown mean by μ and being careful about some other terminology. To do this properly is both subtle and interesting but for another time.

In this case, we presume one knows the distribution. This is fine when we are working with a distribution like geometric or binomial, but with the height and weight example, one may never can truly know the joint distribution. On the other other hand, this frame is a solid starting point and is mathematically rigorous given perfect knowledge of the joint distribution and as the number of samples gets large this knowledge also approaches full information.

5 Linear Regression

5.1 Discussion

We introduced covariance and the correlation coefficient the latter of which is perhaps the most often reported quantity regarding relationships between two quantities. Even in CS70, we often compute this coefficient for midterm scores and final exam scores to "measure" the consistency of our exams. To be sure, this is the primary measure used in science and the first cut in numerous prediction or estimation problems.

For example, you may have heard it said that someone's GPA in their freshman year explains for 80% of the variance in their GPA in their senior year (or something to that effect). This is a mathematical statement about the two random variables freshman year GPA and senior year GPA. The statement is based on a prediction or estimation method called linear regression which we will now discuss.

5.2 The mathematics

We consider two random variables, X and Y on some sample space. Recall that for a given value x for X, the best prediction for Y for minimizing mean squared error is simply $\hat{y}(x) = \mathbb{E}[Y|X=x]$. Note that this prediction could be any function of x defined by the joint distribution.

Instead, we predict the weight using a *linear* function of x. That is, what is the function $\hat{y}(x) = \alpha x + \beta$ that minimizes the mean squared error which is defined as $\mathbb{E}[(Y-\hat{y}(X))^2]$? In particular, we wish to derive values for α and β to produce the best linear estimator for Y given a value x for the random variable X.

For simplicity, we will shift the variables to have mean zero. That is, we consider instead the random variables $\widetilde{X} = X - \mathbb{E}[X]$ and $\widetilde{Y} = Y - \mathbb{E}[Y]$. Note that for a value x for the random variable X, the corresponding value for the random variable $\widetilde{X}$ is $\widetilde{x} = x - \mathbb{E}[X]$.

Furthermore, recall that $\operatorname{Var}(\widetilde{X}) = \operatorname{Var}(X)$, $\operatorname{Var}(\widetilde{Y}) = \operatorname{Var}(Y)$, and $\operatorname{Cov}(\widetilde{X},\widetilde{Y}) = \operatorname{Cov}(X,Y)$ (the variances and covariances aren't affected by shifting the distributions). This follows from the definition of $\operatorname{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$.

To estimate $\widetilde{Y}$ for some value x, we will consider a simpler prediction $\hat{y}(\widetilde{x}) = \alpha\widetilde{x}$ (that is, we assume for now that $b = 0$). We wish to find the value α that minimizes the mean squared error over the entire joint distribution over X and Y, which corresponds to minimizing $\mathbb{E}[(\widetilde{Y} - \alpha\widetilde{X})^2]$.

We can expand this expression as follows:

$$\mathbb{E}[(\widetilde{Y} - \alpha\widetilde{X})^2] = \mathbb{E}[\widetilde{Y}^2] - 2\,\mathbb{E}[\widetilde{X}\widetilde{Y}]\alpha + \mathbb{E}[\widetilde{X}^2]\alpha^2. \tag{2}$$

We get a quadratic function here in α of the form $c + b\alpha + a\alpha^2$, where $a = \mathbb{E}[\widetilde{X}^2]$, $b = -2\,\mathbb{E}[\widetilde{X}\widetilde{Y}]$, and $c = \mathbb{E}[\widetilde{Y}^2]$, which is minimized at $\alpha = -\frac{b}{2a} = \frac{\mathbb{E}[\widetilde{X}\widetilde{Y}]}{\mathbb{E}[\widetilde{X}^2]} = \frac{\operatorname{Cov}(\widetilde{X},\widetilde{Y})}{\operatorname{Var}(\widetilde{X})}$.

We derived this, again, using just our knowledge of a parabola. Again, we assumed that $\beta = 0$ or "goes through" the point $(0,0)$ for $\widetilde{X}$ and $\widetilde{Y}$. We justify this by minimizing $\mathbb{E}[(\widetilde{Y} - (\alpha\widetilde{X} + \beta))^2]$ with respect to the choice of β. Here, we expand to obtain

$$\mathbb{E}[\widetilde{Y}^2] - 2(\alpha\,\mathbb{E}[\widetilde{X}\widetilde{Y}] + \beta\,\mathbb{E}[\widetilde{Y}]) + (\alpha^2\,\mathbb{E}[\widetilde{X}^2] + 2\alpha\beta\,\mathbb{E}[\widetilde{X}] + \beta^2).$$

Collecting terms, taking the derivative with respect to β, and setting the derivative to zero gives the equation:

$$2\,\mathbb{E}[\widetilde{Y}] + 2\alpha\,\mathbb{E}[\widetilde{X}] + 2\beta = 0.$$

Here, we see that $\beta = 0$ satisfies the equation, since $\mathbb{E}[\widetilde{Y}] = \mathbb{E}[\widetilde{X}] = 0$, which implies that this is where the minimum is. So, for the random variables, $\widetilde{X},\widetilde{Y}$ the best linear predictor of $\widetilde{Y}$ is $\hat{y}(\widetilde{x}) = \frac{\operatorname{Cov}(\widetilde{X},\widetilde{Y})}{\operatorname{Var}(\widetilde{X})}\widetilde{x}$.

Finally, to get our prediction $\hat{y}(x)$ of Y, we use the fact that $\operatorname{Cov}(\widetilde{X},\widetilde{Y}) = \operatorname{Cov}(X,Y)$,$\operatorname{Var}(\widetilde{X}) = \operatorname{Var}(X)$, $\operatorname{Var}(\widetilde{Y}) = \operatorname{Var}(Y)$. Substituting yields:

$$\hat{y}(x) = \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}(x - \mathbb{E}[X]) + \mathbb{E}[Y].$$

The intuition above is that if X differs from its expectation, then we vary Y from its expectation by an amount proportional to $\operatorname{Cov}(X,Y)$. The division by $\operatorname{Var}(X)$ scales the movement in X as well as the dependence of the covariance of X on the typical movement in X.

We can also write the formula as follows to see it as a linear function in the variable X.

$$\hat{y}(x) = \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}x - \frac{\operatorname{Cov}(X,Y)}{\operatorname{Var}(X)}\,\mathbb{E}[X] + \mathbb{E}[Y].$$

To get some intuition for why this is the right estimator to use, let's consider two "extreme" cases for X,Y:

- If X and Y are independent, we have $\mathrm{Cov}(X,Y) = 0$, so the best linear predictor is just $\hat{y}(x) = \mathbb{E}[Y]$. This should make sense, since if X,Y are independent, knowing $X = x$ gives us no information about Y, so we might as well use the estimator derived in Section 2.
- If $X = Y$, we have $\mathrm{Cov}(X,Y) = \mathrm{Var}(X)$ and $\mathbb{E}[X] = \mathbb{E}[Y]$. So the best linear predictor is just $\hat{y}(x) = x$. This should also make sense, since if $X = x$ we also know $Y = x$, so $\hat{y}(x) = x$ is the best prediction we can make.

Exercise: Knowing that $\hat{y}(\widetilde{x})$ goes through $(0,0)$ for zero mean random variables X and Y, what point does $\hat{y}(x)$ always go through for general random variables?

5.3 Linear estimation as an explanation of variance

Armed with our derivation, we formally define the phrase that X explains some fraction of the variance of Y by how much smaller the mean squared error for the regression is than the variance of Y (which recall is the mean squared error one gets by using the estimate $\mathbb{E}[Y]$).

Indeed, recall the mean squared error from equation 2 is

$$\mathbb{E}[\widetilde{Y}^2] - 2\alpha\,\mathbb{E}[\widetilde{X}\widetilde{Y}] + \alpha^2\,\mathbb{E}[\widetilde{X}^2],$$

and plugging in $\alpha = \frac{\mathrm{Cov}(\widetilde{X},\widetilde{Y})}{\mathrm{Var}(\widetilde{X})}$, we obtain the expression:

$$\mathbb{E}[\widetilde{Y}^2] - \frac{\mathrm{Cov}(\widetilde{X},\widetilde{Y})^2}{\mathrm{Var}(\widetilde{X})}.$$

Dividing by $\mathrm{Var}(\widetilde{Y}) = \mathbb{E}[\widetilde{Y}^2]$, and recalling that $\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$, we obtain the expression:

$$1 - \mathrm{Corr}(\widetilde{X},\widetilde{Y})^2.$$

We see that the correlation coefficient squared is exactly the fraction by which the mean squared error in the linear regression estimator is less than the error in estimating Y by using the estimate $\mathbb{E}[Y]$. Thus, the square of the correlation coefficient tells us how much of the variance in Y is explained by a linear estimator given X.

6 Statistics

In statistics, one often has some number of samples from a distribution or joint distribution, and one wishes to find the line that “best” fits the data, where the error is defined as the average squared distance to the line. One basic approach is to assume a uniform distribution over the points themselves, and in the limit of large n, the above approach works to find the line.

To be sure, a statistician would find this statement objectionable, as the points only approximate some unknown distribution. They deal with this by thinking through the consequences of the sample mean, variances, and covariances being different from the true ones. Indeed, we have heard exasperation from our statistics colleagues when one confuses the true expectation (typically denoted by μ) with the sample expectation (typically denoted by $\tilde{\mu}$.)

This is beyond this class, but for large data the discussion above is what is happening.

7 Context.

7.1 Other types of error

The mean square error developed here is quite common in science and applications of statistics. You might even read the phrase X “explains variability” in Y in your local newspaper.

There are other measures of error. For example, another error measure is measured as absolute value, that is, we minimize $\mathbb{E}[|X-\hat{x}|]$. In this case, we get that $\hat{x}$ should be a median of the distribution.[3] This particular measure is more "robust" as a far away point has less influence on the value of the median compared to its influence on the value of the expectation.

Other notions of error can be more effective for various applications even while being more challenging to compute. One might see them in future EECS, Data Science, Statistics, or many other departments' curriculums.

7.2 Prediction from more variables

Often one predicts from many variables; e.g., predicting whether an image is a cat or dog from various features in the image. The above discussion can be generalized to estimating a variable Y from many random variables, $X_1,\ldots,X_n$. With a bit of linear algebra, the framework above can be adjusted to this situation. Again, this is an oft-used technique that may be discussed in courses that are in your future.

7.3 Correlation and Causation.

One should be careful. Two random variables may well be correlated (have a positive value for correlation) but one does not "cause" the other; the classic example is that drowning and ice cream consumption are correlated but neither causes the other, the "causal" variable is perhaps summer. In general, causation is a difficult thing to establish. Traditionally, a "gold standard" for establishing causation is considered to be a randomized controlled trial where one randomly assigns treatment or not to understand effect. Indeed, even this process is far from absolute in that various biases can create spurious relationships and often do.

On the other hand, if causation is indeed present one does expect a correlation. Still, even here this is not absolute. For example, the relationship between variables may be non-linear. In health, for example, one sees the "U" shaped curve for various things like weight; being too light or too heavy can be bad for health.

A different aspect of this issue is illustrated by the xor function on two variables. In this case, the output is not correlated to either input, but the two inputs determine the output completely. To be sure, this last example and its extensions is central in theoretical computer science in topics ranging from coding theory to complexity as well as providing statistical lower bounds.

[3] The median of a distribution is a value x where the probability mass above and below x are the same.